LABORATORY OF GENOME INFORMATICS

Assistant Professor
UCHIYAMA, Ikuo

Postdoctoral Fellow: *CHIBA, Hirokazu*

The accumulation of biological data has recently been accelerated by various high-throughput "omics" technologies including genomics, transcriptomics, and proteomics. The field of genome informatics is aimed at utilizing this data, or finding the principles behind the data, for understanding complex living systems by integrating the data with current biological knowledge using various computational techniques. In this laboratory we focus on developing computational methods and tools for comparative genome analysis, which is a useful approach for finding functional or evolutionary clues to interpreting the genomic information of various species. The current focus of our research is on the comparative analysis of microbial genomes, the number of which has been increasing rapidly. Since different types of information about biological functions and evolutionary processes can be extracted from comparisons of genomes at different evolutionary distances, we are developing methods to conduct comparative analyses not only of distantly related but also of closely related genomes.

I. Microbial genome database for comparative analysis

The microbial genome database for comparative analysis (MBGD; URL http://mbgd.genome.ad.jp/) is a comprehensive platform for microbial comparative genomics. The central function of MBGD is to create orthologous groups among multiple genomes using the DomClust program combined with the DomRefine program (see Section II below). By means of these programs, MBGD not only provides comprehensive orthologous groups among the latest genomic data, but also allows users to create their own ortholog groups on the fly using a specified set of organisms. MBGD also has pre-calculated ortholog tables for each major taxonomic group, and provides several views to display the entire picture of each ortholog table. For some closely related taxa, MBGD provides the conserved synteny information calculated using the CoreAligner program. In addition, MBGD provides MyMBGD mode, which allows users to add their own genomes to MBGD. Moreover, MBGD now stores recently accumulating draft genome data, and allows users to incorporate them into a user specific ortholog database through the MyMBGD functionality.

To cope with the rapid growth of microbial genome data, we are trying to establish an efficient protocol to maintain the ortholog tables. We are now developing a progressive strategy to create ortholog data: it first creates an intra-species ortholog table and generates a species pan-genome; next it creates an intra-genera ortholog table and generates a genus pan-genome; and finally it conducts an inter-genera comparison to create an ortholog table covering the entire taxonomic range. This strategy can integrate the multiple ortholog tables currently created: the standard ortholog table created from representative species and the taxon-specific ortholog tables for major taxa. The strategy can also reduce the computation time to calculate all-against-all similarities because it calculates accurate similarities only for inter-genera comparisons and uses a much more rapid (but less accurate) program to calculate intra-species and intra-genus similarities.

II. Orthologous gene classification among multiple genomes at the domain level

As a core technology of our comparative genomics tools, we developed a rapid automated method of ortholog grouping, named DomClust, which is effective enough to allow the comparison of many genomes simultaneously. The method classifies genes based on a hierarchical clustering algorithm using all-against-all similarity data as an input, but it also detects domain fusion or fission events and splits clusters into domains if required. As a result, the procedure can split genes into the domains minimally required for ortholog grouping.

We also developed a procedure to refine the DomClust classification based on multiple sequence alignments instead of pairwise sequence alignments. The method is based on a scoring scheme, the domain-specific sum-of-pairs (DSP) score, which evaluates domain-level classification using the sum total of domain-level alignment scores. We developed a refinement pipeline to improve domain-level clustering, DomRefine, by optimizing the DSP score. DomRefine is now used to construct the standard ortholog table covering all the representative genomes stored in MBGD.

Domain-level classification is a unique feature of our ortholog classification system. In fact, it is different from conventional domain databases like Pfam in that it is based on orthology instead of homology. Particularly, this data is considered to be suitable for analyzing domain fusion events that occurred during evolution. Now, by analyzing the domain-level ortholog grouping data combined with taxonomic and functional information, we are trying to elucidate when and in what kind of genes domain fusion events frequently occurred.

III. Development of a workbench for comparative genomics

We have developed a comparative genomics workbench named RECOG (Research Environment for COmparative Genomics), which aims to extend the current MBGD system by incorporating more advanced analysis functionalities. The central function of RECOG is to display and manipulate a large-scale ortholog table. The ortholog table viewer is a spreadsheet like viewer that can display the entire ortholog table, containing more than a thousand genomes. In RECOG, several basic operations on the ortholog table are defined, which are classified into categories such as filtering, sorting, and coloring, and various comparative analyses can be done by combining these basic operations. In addition, RECOG allows the user to input arbitrary gene properties and compare these properties among orthologs in various genomes.

We continue to develop the system and apply it to various genome comparison studies under collaborative research projects (including *H. pylori* genome comparison described in Section V below). In particular, in addition to microbial genome comparison, we are trying to apply RECOG to comparative analyses of transcriptomic data and metagenomic data.

IV. Ortholog data representation using the Semantic Web technology to integrate various microbial databases

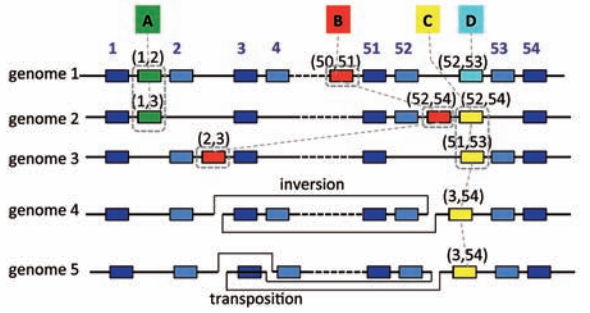
Orthology is a key to integrate knowledge about various organisms through comparative analysis. We have constructed an ortholog database using Semantic Web technology, aiming at the integration of numerous genomic data and various types of biological information. To formalize the structure of the ortholog information in the Semantic Web, we developed an ortholog ontology (OrthO) and described the ortholog information in MBGD in the form of the Resource Description Framework (RDF). On the basis of this framework, we have integrated various kinds of microbial data using the ortholog information as a hub, as part of the MicrobeDB.jp project under the National Bioscience Database Center.

This year, to further standardize the ontology, we developed the Orthology Ontology (ORTH) in collaboration with Dr. Fernandez-Breis (Univ.Murcia) by integrating OrthO and OGO (another ortholog ontology developed by Dr. Fernandez-Breis) and reusing other existing ontologies.

In addition, to facilitate the utilization of the RDF databases distributed worldwide, we developed a command-line tool, named SPANG. SPANG simplifies querying distributed RDF stores using the SPARQL query language, and provides a framework for reusing and sharing queries across the Web, thereby reducing the burden of writing complex queries in SPARQL.

V. *H. pylori* pan-genome analysis for identification of genomic islands

Genomes of bacterial species can show great variation in their gene content, and thus systematic analysis of the entire gene repertoire, termed the "pan-genome", is important for understanding bacterial intra-species diversity. We analyzed the pan-genome identified among 30 strains of the human gastric pathogen *Helicobacter pylori* isolated from various phylogeographical groups. We developed a method (FindMobile) to define mobility of genes against the reference coordinate determined by the core alignment created by CoreAligner, and classified each non-core gene into mobility classes (Figure 1). In addition, by clustering the accessory OGs on the basis of phylogenetic pattern similarity and chromosomal proximity, we identified 60 co-occurring gene clusters (CGCs). Besides known genomic islands including *cag* pathogenicity island, bacteriophages, and integrating conjugative elements, we identified some novel ones, including TerY-phosphorylation triad and that containing a reverse-transcriptase homolog.



OG	mobility extent	non-consecutive	number of genes	N_g	class
A	1	0	2	2	Stable
B	3	0	3	3	Mobile
C	1	2	4	2	Intermediate
D	1	0	1	1	Unique

Figure 1. Definition of the mobility classes in FindMobile. Boxes in dark blue and pale blue represent fully conserved and partially conserved core genes, respectively. Boxes in the other colors are non-core genes that are here classified into four classes: A) stable, B) mobile, C) intermediate, D) unique.

Publication List:

〔Original papers〕

- Fernández-Breis, J.T. Chiba, H., Legaz-García, M.C., and Uchiyama, I. (2016). The orthology ontology: development and applications. J. Biomed. Semant. *7*, 34.
- Hoshino, A., Jayakumar, V., Nitasaka, E., Toyoda, A., Noguchi, H., Itoh, T., Shin-I, T., Minakuchi, Y., Koda, Y., Nagano, A.J., Yasugi, M., Honjo, M.N., Kudoh, H., Seki, M., Kamiya, A., Shiraki, T., Carninci, P., Asamizu, E., Nishide, H., Tanaka, S., Park, K., Morita, Y., Yokoyama, K., Uchiyama, I., Tanaka, Y., Tabata, S., Shinozaki, K., Hayashizaki, Y., Kohara, Y., Suzuki, Y., Sugano, S., Fujiyama, A., Iida, S., and Sakakibara, Y. (2016). Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. Nat. Commun. *7*, 13295.
- Matsui, H., Takahashi, T., Murayama, S., Uchiyama, I., Yamaguchi, K., Shigenobu, S., Suzuki, M., Rimbara, E., Shibayama, K., Overby, A., and Nakamura, M. (2016). Draft genome sequence of *Helicobacter suis* strain SNTW101, isolated from a Japanese patient with nodular gastritis. Genome Announc. *5*, e00934-16.
- Nakai, R., Fujisawa, T., Nakamura, Y., Nishide, H., Uchiyama, I., Baba, T., Toyoda, A., Fujiyama, A., Naganuma, T., and Niki, H. (2016). Complete genome sequence of *Aurantimicrobium minutum* type strain KNCT, a planktonic ultramicrobacterium isolated from river water. Genome Announc. *4*, e00616.
- Uchiyama, I., Albritton, J., Fukuyo, M., Kojima, K., Yahara, K., and Kobayashi, I. (2016). A novel approach to *Helicobacter pylori* pan-genome analysis for identification of genomic islands. PLoS One *11*, e0159419.
- Yahara, K., Furuta, Y., Morimoto, S., Kikutake, C., Komukai, S., Matelska, D., Dunin-Horkawicz, S., Bujnicki, J.M., Uchiyama, I., and Kobayashi, I. (2016). Genome-wide survey of codons under diversifying selection in a highly recombining bacterial species, *Helicobacter pylori*. DNA Res. *23*, 135-143.